

The Bulk and The Tail of Minimal Absent Words in Genome Sequences

Erik Aurell ^{*†}, Nicolas Innocenti ^{*} and Hai-Jun Zhou [‡]

^{*}Department of Computational Biology, KTH Royal Institute of Technology, AlbaNova University Center, SE-10691 Stockholm, Sweden, [†]Department of Information and Computer Science, Aalto University, FI-02150 Espoo, Finland, and [‡]State Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

Minimal absent words (MAW) of a genomic sequence are subsequences that are absent themselves but the subwords of which are all present in the sequence. The characteristic distribution of genomic MAWs as a function of their length has been observed to be qualitatively similar for all living organisms, the bulk being rather short, and only relatively few being long. It has been an open issue whether the reason behind this phenomenon is statistical or reflects a biological mechanism, and what biological information is contained in absent words. In this work we demonstrate that the bulk can be described by a probabilistic model of sampling words from random sequences, while the tail of long MAWs is of biological origin. We introduce the novel concept of a core of a minimal absent word, which are sequences present in the genome and closest to a given MAW. We show that in bacteria and yeast the cores of the longest MAWs, which exist in two or more copies, are located in highly conserved regions the most prominent example being ribosomal RNAs (rRNAs). We also show that while the distribution of the cores of long MAWs is roughly uniform over these genomes on a coarse-grained level, on a more detailed level it is strongly enhanced in 3' untranslated regions (UTRs) and, to a lesser extent, also in 5' UTRs. This indicates that MAWs and associated MAW cores correspond to fine-tuned evolutionary relationships, and suggest that they can be more widely used as markers for genomic complexity.

Minimal absent word; copy-mutation evolution model; random sequence

Abbreviations: AW, absent word; MAW, minimal absent word; rRNA, ribosomal RNA; UTR, untranslated region

Genomic sequences are texts in languages shaped by evolution. The simplest statistical properties of these languages are short-range dependencies, ranging from single-nucleotide frequencies (GC content) to k -step Markov models, both of which are central to gene prediction and many other bioinformatic tasks [1]. More complex characteristics, such as abundances of k -mers, sub-sequences of length k , have applications to classification of genomic sequences [2, 3, 4, 5], and *e.g.* to fast computations of species abundancies in metagenomic data [6, 7, 8, 9].

The reverse image of words present are absent words (AWs), subsequences which actually cannot be found in a text. In genomics the concept was first introduced around 15 years ago for fragment assembly [10, 11] and for species identification [12], and later developed for inter- and intra-species comparisons [13, 14, 15, 16, 17] as well as for phylogeny construction [18]. A practical application is to the design of molecular bar codes such as in the tagRNA-seq protocol recently introduced by us to distinguish primary and processed transcripts in bacteria [19]. Short sequences or tags are ligated to transcript 5' ends, and reads from processed and primary transcripts can be distinguished *in silico* after sequencing based on the tags. For this to be possible it is crucial that the tags do not match any subsequence of the genome under study, *i.e.* that they correspond to absent words. In a further recent study we also showed that the same method allows to separate true antisense transcripts from sequencing artifacts giving a high-fidelity high-throughput antisense transcript discovery

protocol [20]. In these as in other biotechnological applications there is an interest in finding short absent words, preferably additionally with some tolerance.

Minimal absent words (MAW) are absent words which cannot be found by concatenating a substring to another absent word. All the subsequences of a MAW are present in the text. MAWs in genomic sequences have been addressed repeatedly [14, 15, 16, 17, 18] as these obviously form a basis for the derived set of all absent words. Furthermore, while the number of absent words grows exponentially with their length [21], because new AWs can be built by adding letters to other AWs, the number of MAWs for genomes shows a drastically different behavior, as illustrated below in Fig. 1(a) and previously reported in the literature [21, 15, 16]. The behavior can be summarized as there being one or more shortest minimal absent word of a length which we will denote l_0 , a maximum of the distribution at a length we will denote l_{mode} , and a very slow decay of the distribution for large l . In human l_0 is equal to 11, as first found in [13], l_{mode} is equal to 18, there being about 2.25 billion MAWs of that length, while the support of the distribution extends to around 10^6 (Fig. 1(c) and (d)). The total number of human MAWs is about eight billion. As already found in [15] the very end of the distribution depends on the genome assembly; for human Genome assembly GRCh38.p2 the three longest MAWs are 1475836, 831973 and 744232 nt in length. Several aspects of this distribution are interesting. First, in a four-letter alphabet there are 4^k possible subsequences of length k , but in a text of length L only $(L - k)$ subsequences of length k actually appear. If the human genome were a completely random string of letters one would therefore expect the shortest MAW to be of length 15. The fact that l_0 is considerably shorter (11) is therefore already an indication of a systematic bias, in [22] attributed to the hypermutability of CpG sites. We will return to this point below. More intriguing is the observation that the overwhelming majority of the MAWs lie in a smooth distribution around l_{mode} , and then a small minority are found at longer lengths. We will call the first part of the distribution the *bulk* and the second the *tail*. We separate the tail from the bulk by a cut-off l_{max} which we describe below; the human l_{max} is 33, a typical number for larger genomes, while the *Escherichia coli* l_{max} is 24. Using this separation there are about 35 million human tail MAWs, about 0.447% of the total, while there are 7632 *E. coli* tail MAWs, about 0.053% of the total. The effect is qualitatively the same for, as far as we are aware of, all eukaryotic, archeal and bacterial genomes analyzed in the literature [21, 15], as well as all tested by us. Only a few viruses with short genomes are exceptions to this rule and show only the bulk, see Fig. 1(c).

The questions we want to answer in this work are why the distributions of MAWs are described by the bulk and the tail. Can these be understood quantitatively? Do they carry biological information or are they some kind of sampling effects? Can one make further observations? We will show that both the bulk and the tail can be described probabilistically, but in

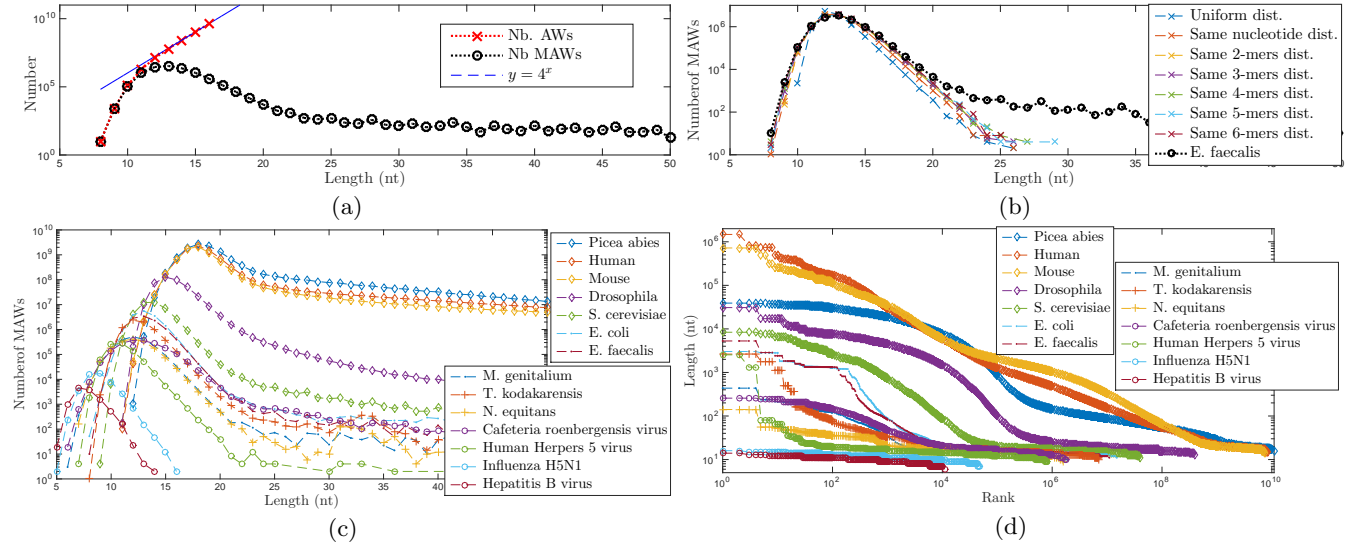


Fig. 1. Distributions of the lengths of absent and minimal absent words in genomes and random texts. (a) Number of AWs and MAWs as a function of word length in the genome of *E. faecalis* v583. The number of AWs grows exponentially while the distribution of MAWs shows a maximum and a decay. (b) Comparison between the distribution of MAWs in *E. faecalis* and the ones for a random genome of the same size using different random models. (c) Distributions of MAWs for a few common organisms and viruses. (d) Lengths of a MAW as a function of its rank for the distributions shown in (c).

two very different ways. The bulk of the MAW distribution arises from sampling words from finite random sequences and are contained in an interval $[l_{\min}, l_{\max}]$, where l_{\max} was introduced above and l_{\min} is a good predictor of l_0 , the actual length of the shortest AWs. To the best of our knowledge this has not been shown previously, and although our analysis uses only elementary considerations, they have to be combined in a somewhat intricate manner. The distribution of bulk MAWs, which comprise the vast majority of MAWs in all genomic sequences, can hence be seen as nothing more than a complicated transformation of simple statistical properties of the sequence. In fact, excellent results are obtained taking only the single nucleotide composition into account (Fig 1(b)). Nevertheless, the tail MAWs are different, and can be described by a statistical model of genome growth by a copy-paste-mutate mechanism similar to the one presented in [23]. We show that the distributions of the tail MAWs vary, both in the data and in the model. The human and the mouse MAW tail distributions follow approximately a power-law, but this seems to be more the exception than the rule; bacteria and yeast as well as *e.g.* *Picea abies* (Norway spruce) show a cross-over behavior to a largest MAW length. For bacteria this largest length ranges from hundreds to thousands; for *P. abies* it is around 30 000; while for human and mouse the tail MAW distribution reaches up to one million, without cross-over behavior (Fig 1(d)).

From the definition, any subword obtained by removing letters from the start or end of a MAW is present in the sequence. In particular, removing the first and last letters of a MAW leads to a subword that is present at least twice, which we denote here as a *MAW core*. MAWs made of a repeat of the same letter are an exception to this rule as they can have the two copies of their cores overlapping each other, see Appendix A in *Supplemental Information*. MAW cores can be considered as the causes that create the MAWs and their location on the genome combined with functional information from the annotation tells us about their biological significance. Finding all the occurrences on a genome of a given word (such as a MAW core) is the very common bioinformatic task of alignment, which can be done quickly and efficiently using one of the many software packages available. In bacteria and yeast,

the cores from the longest MAWs are predominantly found in regions coding for ribosomal RNAs (rRNAs), regions present in multiple copies on the genome and under high evolutionary pressure as their sequence determines their enzymatic properties, required for protein synthesis and vital to every living cell. At the global scale, it appears that MAW cores obtained from MAWs in the bulk are distributed roughly uniformly over the genome while those from the tail cluster in 3' UTRs and, to a lesser extent, also in 5' UTRs. These regions are important for post-transcriptional regulation, and thus likely to be under evolutionary pressure similarly to rRNAs.

We end this Introduction by noting that from a linguistic perspective a language can be described by its list of *forbidden* sub-sequences, or the list of its forbidden words [24]¹. Minimal forbidden words relate to forbidden words as MAWs to absent words, and in a text of infinite length the lists of MAWs and minimal forbidden words would agree. If there is a finite list of minimal forbidden words the resulting language lies on the lowest level of regular languages in the Chomsky hierarchy [25, 26], and is hence relatively simple, while a complex set of instructions, such as a genome, is expected to correspond to a more complex language, with many layers of meaning. Such aspects have been exploited in cellular automata theory [27, 28] and in dynamical systems theory [29], and are perhaps relevant to genomics as well. The present investigation is however focused on properties of texts of finite length, for which minimal forbidden words and MAWs are quite different.

A random model for the bulk

Let us consider a random sequence \mathcal{S} of total length N with alphabet $\{A, C, G, T\}$. Each position of \mathcal{S} is independently assigned the letter A , C , G , or T with corresponding probabilities ω_A , ω_C , ω_G and ω_T ($\equiv 1 - \omega_A - \omega_C - \omega_G$). A word of

¹In genomics the concept of minimal absent words was first introduced in [11], as “minimal forbidden words”. Since this term has another well-established meaning we have here instead used MAW [21]. Related concepts are “unwords” [14] which are the shortest absent words (also shortest minimal absent words) and “nullomers” [13] which are absent words without a requirement on minimality, compare data in Table 1 in [13].

length L has the generic form of $\mathbf{w} \equiv c_1 c_2 \dots c_{L-1} c_L$, where $c_i \in \{A, C, G, T\}$ is the letter at the i -th position. The total number of such words is 4^L . This number exceeds N when L increases to order $O(\ln N)$, therefore most of the words of length $L \geq O(\ln N)$ will never appear in \mathcal{S} . Then what is the probability $q_{\mathbf{w}}$ of a particular word \mathbf{w} being a MAW of sequence \mathcal{S} ?

For \mathbf{w} to be a MAW, it must not appear in \mathcal{S} but its two subwords of length $(L-1)$, $\mathbf{w}^{(p)} \equiv c_1 c_2 \dots c_{L-1}$ and $\mathbf{w}^{(s)} \equiv c_2 \dots c_{L-1} c_L$, must appear in \mathcal{S} at least once, as demonstrated schematically in Fig. 2(a). We define the core of the MAW \mathbf{w} as the substring

$$\mathbf{w}^{\text{core}} \equiv c_2 \dots c_{L-1},$$

which must appear in \mathcal{S} at least twice, except for the special case of $c_1 = c_2 = \dots = c_L$ where the $\mathbf{w}^{(p)}$ and $\mathbf{w}^{(s)}$ overlap (see Appendix A in *Supplemental Information*). The core must immediately follow the letter c_1 at least once and it must also be immediately followed by the letter c_L at least once. Similarly, if \mathbf{w}^{core} immediately follows the letter c_1 , it must not be immediately followed by the letter c_L .

We can construct $(N-L+1)$ subsequences of length L from \mathcal{S} , say $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{N-L+1}$. Neighboring subsequences are not fully independent as there is an overlap of length $(L-m)$ between \mathcal{S}_n and \mathcal{S}_{n+m} with $1 \leq m < L$. However, for $L \ll N$ two randomly chosen subsequences of length L from the random sequence \mathcal{S} have a high probability of being completely uncorrelated. We can thus safely neglect these short-range correlations, and consequentially the probability of word \mathbf{w} being a MAW is expressed as

$$q_{\mathbf{w}} = [1 - \omega(\mathbf{w})]^{N-L+1} - \left\{ [1 - \omega(\mathbf{w}^{(p)})]^{N-L+2} + [1 - \omega(\mathbf{w}^{(s)})]^{N-L+2} - [1 - \omega(\mathbf{w}^{(p)}) - \omega(\mathbf{w}^{(s)}) + \omega(\mathbf{w})]^{N-L+1} \right\}, \quad [1]$$

where $\omega(\mathbf{w}) \equiv \prod_{i=1}^L \omega_{c_i}$ (with c_i being the i -th letter of \mathbf{w}) is the probability of a randomly chosen subsequence of length L from \mathcal{S} to be identical to the word \mathbf{w} , while $\omega(\mathbf{w}^{(p)})$ and $\omega(\mathbf{w}^{(s)})$ are, respectively, the probabilities of a randomly chosen subsequences of length $(L-1)$ from \mathcal{S} being identical to $\mathbf{w}^{(p)}$ and $\mathbf{w}^{(s)}$.

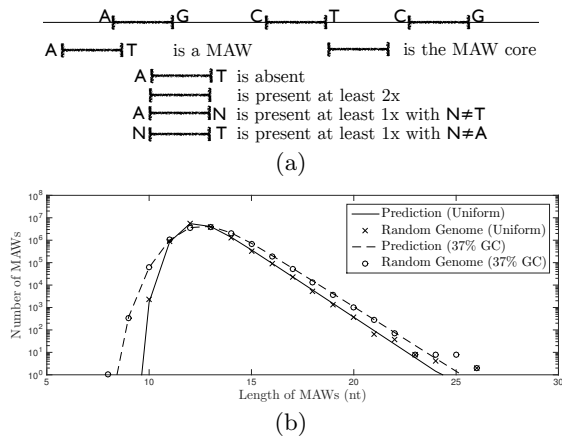


Fig. 2. (a) Illustration of the properties of a minimal absent word and its subwords. (b) Comparison between the length distribution predicted by Eq. 2 and the number of MAWs calculated for one instance of a random genome of 3.3 Mbp with uniform nucleotide distribution and with 37% GC content.

Summing over all the 4^L possible words of length L , we obtain the expected number $\bar{\Omega}(L) \equiv \sum_{\mathbf{w}} q_{\mathbf{w}}$ of MAWs of length L for a random sequence \mathcal{S} of length N :

$$\begin{aligned} \bar{\Omega}(L) = & \sum_{c_1} \sum_{c_L} \sum_{n_A, n_C, n_G, n_T} \frac{(L-2)!}{n_A! n_C! n_G! n_T!} \delta_{n_A + n_C + n_G + n_T}^{L-2} \\ & \times \left\{ (1 - \omega_{c_1} \omega_{c_L} \omega_A^{n_A} \omega_C^{n_C} \omega_G^{n_G} \omega_T^{n_T})^{N-L+1} \right. \\ & - (1 - \omega_{c_1} \omega_A^{n_A} \omega_C^{n_C} \omega_G^{n_G} \omega_T^{n_T})^{N-L+1} \\ & - (1 - \omega_{c_L} \omega_A^{n_A} \omega_C^{n_C} \omega_G^{n_G} \omega_T^{n_T})^{N-L+1} \\ & \left. + (1 - (\omega_{c_1} + \omega_{c_L} - \omega_{c_1} \omega_{c_L}) \omega_A^{n_A} \omega_C^{n_C} \omega_G^{n_G} \omega_T^{n_T})^{N-L+1} \right\}, \quad [2] \end{aligned}$$

where the summation is over all the 16 combinations of the two terminal letters c_1, c_L and over all the possibilities with which the letters A, C, G, T , may appear in the core a total number of times equal to respectively n_A, n_C, n_G , and n_T .

In the simplest case of maximally random sequences, namely $\omega_A = \omega_C = \omega_G = \omega_T = \frac{1}{4}$, Eq. (2) reduces to

$$\begin{aligned} \bar{\Omega}(L) = & 4^L (1 - 4^{-L})^{N-L+1} \\ & \times \left[1 - 2 \left(1 - \frac{3}{4L-1}\right)^{N-L+1} + \left(1 - \frac{6}{4L-1}\right)^{N-L+1} \right]. \quad [3] \end{aligned}$$

We have checked by numerical simulations (see Fig. 2b) that Eq. 2 and Eq. 3 indeed give excellent predictions of the number MAWs as a function of their length in random sequences.

We define a predicted minimum and a predicted maximum of the support of the bulk (l_{\min} and l_{\max}) as the two values of L such that $\bar{\Omega}(L) = 1$. In the general case, requiring that $\bar{\Omega}(L) \geq 1$ we obtain $L \geq l_{\min}$, with the shortest length l_{\min} such that

$$\sum_{n_A, n_C, n_G, n_T} \frac{l_{\min}!}{n_A! n_C! n_G! n_T!} \delta_{n_A + n_C + n_G + n_T}^{l_{\min}} \times (1 - \omega_A^{n_A} \omega_C^{n_C} \omega_G^{n_G} \omega_T^{n_T})^{N-l_{\min}} = 1. \quad [4]$$

is closest to one, while in the other limit we obtain that $L \leq l_{\max}$, with the longest length l_{\max} being

$$l_{\max} \approx \frac{2 \ln N}{-\ln(\omega_A^2 + \omega_C^2 + \omega_G^2 + \omega_T^2)}. \quad [5]$$

The bulk distribution is therefore centered around lengths of order $\log N$. In the case of maximally random sequences, we can obtain the lower limit analytically and also the first correction to (5), as

$$l_{\min} \simeq \frac{\ln N - \ln \ln N}{\ln 4}, \quad [6]$$

and

$$l_{\max} \simeq \frac{2 \ln N + \ln 9}{\ln 4}. \quad [7]$$

The above definition of l_{\max} is good enough for our purposes, and l_{\min} is also a good predictor for l_0 (see below). A more refined predictor for l_{\min} is discussed in Appendix B in the *Supplemental Information*.

A random model for the tail

We now describe a protocol for constructing random genome by a iterative copy-paste-mutation scheme that qualitatively reproduces the tail behavior observed for most of real genomes.

The model is in principle similar to [23] but differs in the details of the implementation.

The starting point is a string of nucleotides chosen independently at random with a length N_0 . At each iteration, we chose two positions i and j uniformly at random on the genome and a length l from a Poisson distribution with mean λ . We copy the sequence between i to $(i + l - 1)$ and insert it between positions j and $j + 1$, thus increasing the genome size by l . We then randomly alter a fraction α of nucleotides in the genome, choosing the positions uniformly at random and the new letters from an arbitrary distribution that can be tuned to adjust the GC content. The process is repeated until the genome reaches the desired length.

In this model, λ represents the typical size of region involved in a translocation in the genome and α corresponds to the expected number of mutations between such events. We observed that the length N_0 or the content of the initial string is unimportant provided that it is much shorter than the final genome size. The exact value of λ given a constant α/λ ratio only affects the tail of the distribution far away from the bulk. We have also checked by simulations that using different distributions for the choice of l does not affect the results qualitatively (See Fig. S1 in *Supplemental Information*).

A low ratio α/λ generates genomes with tail MAWs while higher values cause them to only have bulk MAWs as in random texts discussed above (Fig. 3). This is in agreement with the observations for viruses in Fig. 1: *Hepatitis B* and *H5N1* are viruses that replicate using an error prone reverse transcription and the MAW distributions for their genomes lack the tail. In contrast, *Human Herpes 5* virus and the *Cafeteria roenbergensis* virus are DNA viruses that use the higher fidelity DNA replication mechanism and their genomes clearly have tail MAWs.

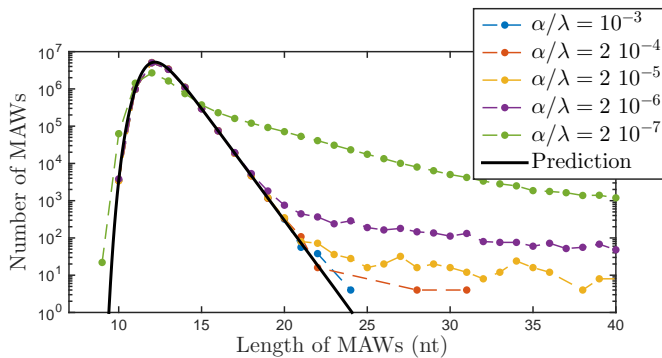


Fig. 3. Length distributions for MAWs in a few random genomes 3 millions bp in size generated by the copy-paste-mutation protocol with different values for the α/λ ratio. The curves were obtained using $N_0 = 5000$ and a constant $\lambda = 500$.

Estimating the length of the shortest absent words

Equations (6) and (4) can be used to estimate the length of the shortest absent words. Fig. 4 compares the prediction of the simplest estimate in Eq. 6 to the length of the shortest MAW for a large set of genomes.

The estimator Eq. 6 is expected to be most accurate only for genomes with neutral GC content. The figure reveals that genomes of comparable sizes typically vary in their l_0 by about 4 nt, and that our estimator captures very well the upper values in this distribution. Using Eq. 4 only improves the predictions for genomes with much biased GC content (40% or

less) and leads to results in line with the earlier published estimator by Wu et al. [17], see Appendix C and Table S1 in *Supplemental Information*.

This analysis shows that, contrary to the conclusion of [22], there is no need to invoke a biological mechanism to explain the length of the shortest MAW; it is instead a property of rare events when sampling from a random distribution. Indeed, the estimator Eq. 4 gives the length of the shortest Human MAW as 12 and not 15, only one nucleotide away from the correct answer (11).

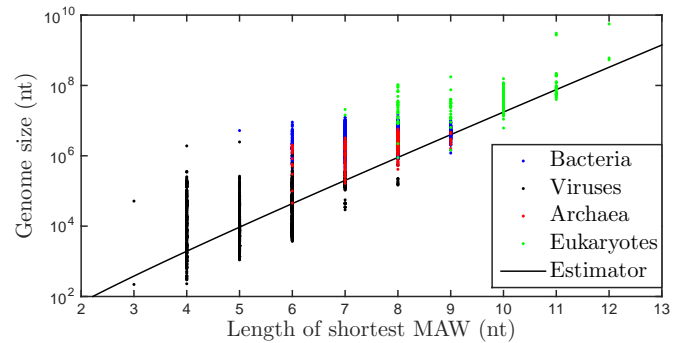


Fig. 4. Length of the shortest MAW versus the genomes size for all viral, bacterial and archaeal genomes available on the NCBI database as well as a few arbitrarily chosen eukaryotes, including many short genomes and only a few complex organisms, such as Human, mouse and Norway spruce. The black line represents the estimator Eq.6.

The origins of tail MAWs

The Human Herpes virus 5, a double stranded DNA virus with a linear genome of ~ 235.6 kbp, has four very long MAWs with lengths of 2540 nt for two of them, and 1360 nt for two others, all other MAWs being much shorter (81 nt or less). The cores of these four MAWs come from three regions, two of them located at the very beginning and very end of the genome, and the third at position ~ 195 kbp made up of the juxtaposition of the reverse complements of the two others. These regions are annotated as repeated and regulatory. Based on the NCBI BLAST webservice, these particular sequences are highly conserved (95% or more) in numerous strains of the virus and do not seem to have homologues in any other species: the closely related *Human Herpes virus 2* shows sequences with no more than 42% similarities to these MAW cores.

In *E. coli* and *E. faecalis*, the 10 longest MAWs with lengths between 2815 and 3029 nt all originate from rRNA regions: a set of genes present in a few copies made of highly conserved regions with minor variations between the copies. For yeast, the four longest MAWs (8376 nt) originate from the two copies of rRNA RDN37 on chromosome XII, another four (7620 nt) are caused by 2 copies of the region containing PAU1 to VTH2 on chromosome X and PAU4 to VTH2 on chromosome IX and two more (6531 nt) originate from the copies of gene YRF on chromosome VII and XVI (YRF is present in at least 8 copies on the yeast genome).

We performed an extensive search for all occurrences of the cores of every MAW found in the organisms mentioned above and considered the density of MAW cores along the genome for MAWs in the tail (see Fig. S2 in *Supplemental Information*). Except for the Human Herpes virus 5 that only has few MAWs which cluster in the repeated segments discussed above, MAW cores do not appear to be preferentially located in any specific regions on the genome scale. A more detailed analysis

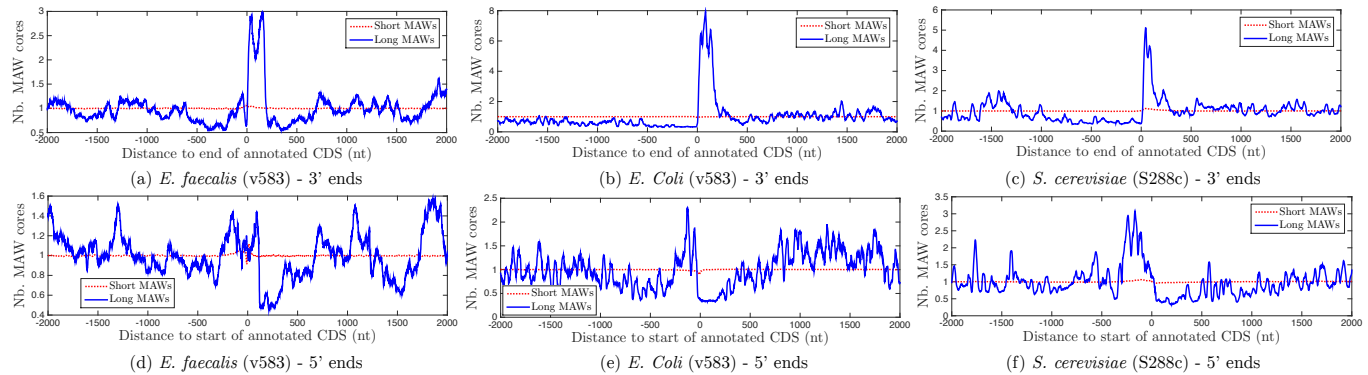


Fig. 5. Average number of MAW cores for MAWs from the bulk and the tail around (a)-(c) ends (3' UTRs) and (d)-(f) starts (5' UTRs) of annotated CDS for three living organisms (see also Fig. S2 in *Supplemental Information*). The signals are normalized so that their mean value over the window of interest is 1. The MAW cores concentrate in the UTRs, more strongly on the 3' side than on the 5' side.

(Fig. 5) however reveals that, while cores of MAWs from the bulk appear uniformly distributed, those from the tail cluster downstream of ends of annotated coding DNA sequences (CDS) (i.e. in the 3' UTRs and terminator sequences). A similar yet weaker effect can be observed upstream of the start of annotated CDS (the 5' UTR). By definition, a MAW core corresponds to a repeated region on the genome immediately surrounded by nucleotides varying between the copies. Exact repeated regions lead to only a few MAWs with cores corresponding to that repeat. Introducing a few random changes in such regions creates more but shorter MAWs, the cores of which are the sub-strings common to two or more regions. A high density of MAW cores in a family of regions such as the UTRs thus indicates that they share a limited set of building blocks, implying a similar set of evolutionary constraints or a common origin.

The significance of the longest MAWs

We now consider the lengths of the longest MAW found in the genomes of numerous organisms and viruses. We observe that this length generally lies between l_{\max} and a length of about 10% of the genome size (Fig. 6).

Viruses are the class showing the largest spread in the length of their longest MAWs. Many viruses are close to l_{\max} , particularly for those with shorter genomes, confirming our previously mentioned observation that some lack the tail and are thus closer to random texts. Nevertheless, a few viral genomes have MAWs longer than 10% of the genome length, i.e. proportionally longer than in any living organism. The figure suggests that bacteria have on average slightly longer MAWs than archaea, but overall no clear distinctions between the four types of genomes can be noted based on the length of MAWs alone, suggesting that the mechanisms behind evolution of all organisms and viruses influence the MAWs distribution in the same way.

A more detailed analysis of the data presented in Fig. 5 shows that organisms have the longest MAWs closest to the lower bound of the tail (l_{\max}) or the observed upper bound of 10% of the genome length share some common traits.

In bacteria, the 6 genomes having their longest MAWs closest to l_{\max} are two strains from the *Buchnera aphidicola* species, two strains of *Candidatus Carsonella ruddii*, one *Candidatus Phytoplasma solani* and one *Bacteroides uniformis*. While the last is a putative bacterial species living in human feces [30], the five other species are intra-cellular symbiotic or parasitic gammaproteobacteria in plants or insects [31, 32, 33].

Among eukaryotes, the same analysis gives us *Plasmodium gaboni*, an agent responsible for malaria [34], a species of *Cryptosporidium*, another family of intracellular parasites found in drinking water and *Chromera velia*, a photosynthetic organism from the same apicomplexa phylum as plasmodium, which is remarkable in this class for its ability to survive outside a host and is of particular interest for studying the origin of photosynthesis in eukaryotes [35, 36]. For Archaea, we find *Candidatus Parvarchaeum acidiphilum* and *C. P. acidophilus*, which are two organisms with short genomes (45.3 and 100 kbp) living in low pH drainage water from the Richmond Mine in Northern California [37], and an uncultivated hyperthermophilic archaea "SCGC AAA471-E16" of the Aigarchaeota phylum [38]. Additionally, we searched for MAWs in 2395 human mitochondrial genomes with lengths between 15436 and 16579 bp and found that the longest MAWs are only 17 or 18 nt long, while $l_{\max} \simeq 16.5$ for these genomes.

Among bacteria having their longest MAW close to 10% of the genome length, we find several strains of *E. coli*, *Francisella tularensis*, *Shewanella baltica*, *Methylobacillus flagellatus*, *Xanthomonas oryzae* and a species of the *Wolbachia* genus. All of these are facultative or obligatory aerobes and are a lot more widespread than the bacteria listed in the previous paragraph. At least the four first species are free-living and commonly cultured in labs. *X. oryzae* is a pathogen affecting rice residing in the intercellular spaces. *Wolbachia* species are a very common intracellular parasites or symbiotes living in arthropods. Among eukaryotes, in addition to human and

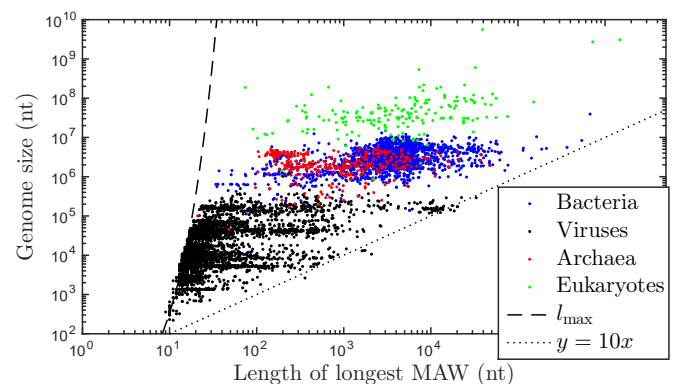


Fig. 6. Length of the longest absent words as a function of the genome sizes for the same set of genomes as in Fig. 4. The dashed line represents the estimator in Eq. 7 and the dotted line $y = 0.1x$.

mouse, we find *Dictyostelium discoideum*, an organism living in soil that changes from uni- to multi-cellular during its life cycle, and *Thalassiosira Pseudonana* a unicellular alga commonly found in marine phytoplankton. Finally, archaea with the longest MAWs are *Methanococcus voltae*, a mesophilic methanogen, *Halobacterium salinarum*, a halophilic marine obligate aerobic archeon also found in food such as salted pork or sausages, and *Halalkalicoccus jeotgali*, another halophilic archeon isolated from salted sea-food[39].

To summarize, it appears from these results that development in specific environmental niches that offer rather stable conditions, and particularly the inside of the cell of another organism, is, albeit with some exceptions, associated with short MAWs. On the other hand, aerobic life and widespread presence in changing and diverse environments such as soil, sea water or food is generally associated with long MAWs.

Intracellular organisms have a well known tendency to reduce the size of their genomes and increase error rate for DNA replication due to elimination of error correction mechanisms [40]. This translates into a high value for α in our random model for the tail and explains why their longest MAWs are short. In particular, we note that *B. aphidicola* (genome size of ~ 650 kbp), brought out by our analysis, is known to have the highest mutation rate among all prokaryotes and indeed its longest MAW is as short as 34 nt [40, 41]. As for organisms specialized for a niche environment, one may hypothesize that proliferation speed is more important than replication fidelity [41].

Reasons for a widespread presence in changing environments to increase the length of the longest MAWs are less clear. Multicellular eukaryotic organisms do not show particularly high DNA replication fidelity [41]. The fidelity of *E. coli* is fairly good at about 3.5x the one of human germline [41], but it is unclear if this is enough to make it stand out from other bacteria. We speculate that soil, sea water or food, which are likely to contain many types of microorganisms, may favor species more likely to undergo horizontal gene transfers, thus increasing the rate of translocation events and decreasing α in our model, while leaving the *per generation* mutation rate unchanged.

Conclusions

We have proposed a two-parts model explaining the unusual shape of the length distribution of minimal absent words (MAWs) in genomes. The first part of the model quantitatively reproduces the bulk of the distribution by considering the genome as a random text with random and independent letters. The second part is a stochastic algorithm grounded in basic principles of how genomes evolve, through translocation events and mutations, that qualitatively reproduces the behavior in the tail. Our theory provides an estimator for the

length of the shortest MAWs that is remarkably simple and captures well the global trend observed in large numbers of genomes from all sorts of organisms and viruses.

Considerations about the longest MAW in a genome reveal sets of organisms sharing common high-level features such as the type of environment they live in. We have shown arguments for believing that organisms and viruses having few tail MAWs do so because of a low replication fidelity. Why some organisms such as *E. coli* have long tail MAWs is less clear and replication fidelity alone does not seem to be a sufficient reason.

Finally, we have introduced the concept of MAW cores, sequences present on the genome that tell us about what causes the existence of their parent MAW. We have shown that, while cores from bulk MAWs do not seem biologically relevant, cores from tail MAWs cluster in regions of the genome with special roles, namely ribosomal RNAs and untranslated regions surrounding coding regions of genes, a feature that cannot be explained by a stochastic protocol that ignores the biological roles of the strings it manipulates.

Materials and Methods

Data source. Viral and bacterial genomes were downloaded under the form of the "all.fna" archives from the "Genomes/Viruses/" and "Genomes/Bacteria" from the NCBI database on 17-18 May 2015 respectively. The Norway spruce's genome was downloaded from the "Spruce genome project" [42] homepage and the yeast genome strain 288C [43] from Saccharomyces Genome Database. Genomes of other eukaryotes and archaeas were downloaded from the NCBI database at several different dates over the period May-June 2015. The human mitochondrial genomes were downloaded from the Human Mitochondrial Genome Database (mtDB)[44] in early September 2015.

Software & Computational Ressources. All MAWs were computed using the software provided by Pinho et al. in [21]. The software was run taking into account the reverse-complementary strand ('-rc' command line switch) and requesting MAWs ('-n' command line option) with length up to five million nucleotides, i.e. much longer than the expected length of the longest MAWs. The search for MAWs was performed on commodity desktop computers for all but the Human, mouse, and Norway spruce genomes, for which the computer "Ellen"² from the Center for High Performance Computing (PDC) at KTH was used. Localization of occurrences of MAW cores on the genomes for figures 5 and S2 was done by aligning these subwords to their respective genomes using Bowtie2 [45], allowing only strict alignments (command line option '-v 0'). Identical cores from different MAWs are counted independently in the coverage.

ACKNOWLEDGMENTS. This research is supported by the Swedish Science Council through grant 621-2012-2982 (EA), by the Academy of Finland through its Center of Excellence COIN (EA), and by the Natural Science Foundation of China through grant 11225526 (HJZ). EA thanks the hospitality of KITPC and HJZ thanks the hospitality of KTH. The authors thank Profs Rüdiger Urbanke and Nicolas Macris, and Dr. Françoise Wessner for valuable discussions, and PDC, the Center for High Performance Computing at KTH, for access to computational resources needed to analyze large genomes.

1. Durbin, R, Eddy, S. R, Krogh, A, & Mitchison, G. (1998) Biological sequence analysis. (Cambridge University Press).
2. Sandberg, R, Brändén, C.-I, Ernberg, I, & Cöster, J. (2003) Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. *Gene* 311, 35–42.
3. Hao, B.-L & Qi, J. (2004) Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance. *J. Bioinformatics and Comput. Biol.* 2, 1–19.
4. Chor, B, Horn, D, Goldman, N, Levy, Y, & Masingham, T. (2009) Genomic DNA k-mer spectra: models and modalities. *Genome Biology* 10.
5. Rosa, M.-L, Fiannaca, A, Rizzo, R, & Urso, A. (2015) Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC Bioinformatics* 16.
6. Amir, A & Zuk, O. (2011) Bacterial community reconstruction using compressed sensing. *Journal of Computational Biology* 18, 1723–1741.
7. Amir, A, Zeisel, A, Zuk, O, Elgart, M, Stern, S, Shamir, O, Turnbaugh, J, Soen, Y, & Shental, N. (2013) High-resolution microbial community reconstruction by integrating short reads from multiple 16S rRNA regions. *Nucleic Acids Res.* 41.
8. Koslicki, D, Foucart, S, & Rosen, G. (2013) Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics* 19, 2096–2102.
9. Chatterjee, S, Koslicki, D, Dong, S, Innocenti, N, Cheng, L, Lan, Y, Vehkaperä, M, Skoglund, M, Rasmussen, L. K, Aurell, E, & Corander, J. (2014) SEK: sparsity exploiting k-mer-based estimation of bacterial community composition. *Bioinformatics* 30, 2423–2431.
10. Mignosi, F, Restivo, A, & Sciortino, M. (2001) Forbidden factors and fragment assembly. *RAIRO Theoret. Inform. Appl.* 35, 565–577.

²<https://www.pdc.kth.se/resources/computers/ellen>

11. Fici, G, Mignosi, F, Restivo, A, & Sciortino, M. (2006) Word assembly through minimal forbidden words. *Theor. Comput. Sci.* 359, 214–230.
12. Fofanov, Y, Fofanov, Y, & Pettitt, B. (2002) Counting array algorithms for the problem of finding appearances of all possible patterns of size n in a sequence. (W.M. Keck Center for Computational and Structural Biology).
13. Hampikian, G & Andersen, T. (2007) Absent sequences: nullomers and primes. *Pac Symp Biocomput* pp. 355–366.
14. Herold, J, Kurtz, S, & Giegerich, R. (2008) Efficient computation of absent words in genomic sequences. *BMC Bioinformatics* 9, 167.
15. Garcia, S, P, Pinho, A, J, Rodrigues, J, M. O. S, Bastos, C. A. C, & Ferreira, P. J. S. G. (2011) Minimal absent words in prokaryotic and eukaryotic genomes. *PLoS ONE* 6, e16065.
16. Garcia, S. P & Pinho, A. J. (2011) Minimal absent words in four human genome assemblies. *PLoS ONE* 6, e29344.
17. Wu, Z.-D, Jiang, T, & Su, W.-J. (2010) Efficient computation of shortest absent words in a genomic sequence. *Information Processing Letters* 110, 596 – 601.
18. Chairungsee, S & Crochemore, M. (2012) Using minimal absent words to build phylogeny. *Theoretical Computer Science* 450, 109 – 116. Implementation and Application of Automata (CIAA 2011).
19. Innocenti, N, Golumbeanu, M, d'Hérouël, A. F, Lacoux, C, Bonnin, R. A, Kennedy, S. P, Wessner, F, Serror, P, Boulou, P, Repoila, F, & Aurell, E. (2015) Whole genome mapping of 5' ends in bacteria by tagged sequencing: A comprehensive view in *Enterococcus faecalis*. *RNA* 21, 1018–1030.
20. Innocenti, N, Repoila, F, & Aurell, E. (2015) Detection and quantitative estimation of spurious double stranded dna formation during reverse transcription in tagma-seq. *RNA Biology*.
21. Pinho, A, Ferreira, P, Garcia, S, & Rodrigues, J. (2009) On finding minimal absent words. *BMC Bioinformatics* 10, 137.
22. Acquisti, C, Poste, G, Curtiss, D, & Kumar, S. (2007) Nullomers: Really a matter of natural selection? *PLoS ONE* 2, e1022.
23. Hsieh, L.-C, Luo, L, Ji, F, & Lee, H. C. (2003) Minimal model for genome evolution and growth. *Phys. Rev. Lett.* 90, 018101.
24. Hopcroft, E & Ullman, J. (1979) *Introduction to Automata Theory, Languages and Computation*. (Addison-Wesley).
25. Chomsky, N. (1956) Three models for the description of language. *IRE Transactions on Information Theory* 2, 113–124.
26. Chomsky, N & Schützenberger, M. P. (1963) *The Algebraic Theory of Context-Free Languages*. (North-Holland, Amsterdam), pp. 118–161.
27. Wolfram, S. (1984) Computation theory of cellular automata. *Communications in Mathematical Physics* 96, 15–57.
28. Nordahl, M. (1989) Formal languages and finite cellular automata. *Complex Systems* 3, 63–78.
29. Auerbach, D, Cvitanovic, P, Eckmann, J, Gunaratne, G, & Procaccia, I. (1987) Exploring chaotic motion through periodic orbits. *Physical Review Letters* 58, 2387–2389.
30. Renouf, M & Hendrich, S. (2011) *Bacteroides uniformis* is a putative bacterial species associated with the degradation of the isoflavone genistein in human feces. *The Journal of Nutrition* 141, 1120–1126.
31. Quaglino, F, Zhao, Y, Casati, P, Bulgari, D, Bianco, P. A, Wei, W, & Davis, R. E. (2013) 'Candidatus *Phytoplasma solani*', a novel taxon associated with stolbur and bois noir related diseases of plants. *International journal of systematic and evolutionary microbiology* pp. ijs–0.
32. Tamames, J, Gil, R, Latorre, A, Pereto, J, Silva, F, & Moya, A. (2007) The frontier between cell and organelle: genome analysis of *Candidatus Carsonella ruddii*. *BMC Evolutionary Biology* 7, 181.
33. van Ham, R. C. H. J, Kamerbeek, J, Palacios, C, Rausell, C, Abascal, F, Bastolla, U, Fernández, J. M, Jiménez, L, Postigo, M, Silva, F. J, Tamames, J, Viguera, E, Latorre, A, Valencia, A, Morán, F, & Moya, A. (2003) Reductive genome evolution in *Buchnera aphidicola*. *Proceedings of the National Academy of Sciences* 100, 581–586.
34. Ollomo, B, Durand, P, Prugnolle, F, Douzery, E, Arnathau, C, Nkoghe, D, Leroy, E, & Renaud, F. (2009) A new malaria agent in african hominids. *PLoS Pathog* 5, e1000446.
35. Keeling, P. J. (2008) Evolutionary biology: bridge over troublesome plastids. *Nature* 451, 896–897.
36. Moore, R. B, Oborník, M, Janoušková, J, Chrudimský, T, Vancová, M, Green, D. H, Wright, S. W, Davies, N. W, Bolch, C. J, Heimann, K, et al. (2008) A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* 451, 959–963.
37. Baker, B. J, Comolli, L. R, Dick, G. J, Hauser, L. J, Hyatt, D, Dill, B. D, Land, M. L, VerBerkmoes, N. C, Hettich, R. L, & Banfield, J. F. (2010) Enigmatic, ultra-small, uncultivated archaea. *Proceedings of the National Academy of Sciences* 107, 8806–8811.
38. Rinke, C, Schwientek, P, Sczyrba, A, Ivanova, N. N, Anderson, I. J, Cheng, J.-F, Darling, A, Malfatti, S, Swan, B. K, Gies, E. A, Dodsworth, J. A, Hedlund, B. P, Tsiamis, G, Sievert, S. M, Liu, W.-T, Eisen, J. A, Hallam, S. J, Kyrpides, N. C, Stepanauskas, R, Rubin, E. M, Hugenholtz, P, & Woyke, T. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437.
39. Roh, S. W, Nam, Y.-D, Chang, H.-W, Sung, Y, Kim, K.-H, Oh, H.-M, & Bae, J.-W. (2007) *Halalkalicoccus jeotgali* sp. nov., a halophilic archaeon from shrimp jeotgal, a traditional korean fermented seafood. *International Journal of Systematic and Evolutionary Microbiology* 57, 2296–2298.
40. Moran, N. A, McLaughlin, H. J, & Sorek, R. (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323, 379–382.
41. Lynch, M. (2010) Evolution of the mutation rate. *Trends in genetics* : TIG 26, 345–352.
42. Nystedt, B, Street, N. R, Wetterbom, A, Zuccolo, A, Lin, Y.-C, Scofield, D. G, Vezzi, F, Delhomme, N, Giacomello, S, Alexeyenko, A, Vicedomini, R, Sahlin, K, Sherwood, E, Elfstrand, M, Gramzow, L, Holmberg, K, Hallman, J, Keech, O, Klasson, L, Koriabine, M, Kucukoglu, M, Kaller, M, Luthman, J, Lysholm, F, Niittyla, T, Olson, A, Rilakovic, N, Ritland, C, Rossello, J. A, Sena, J, Svensson, T, Talavera-Lopez, C, Theissen, G, Tuominen, H, Vanneste, K, Wu, Z.-Q, Zhang, B, Zerbe, P, Arvestad, L, Bhalerao, R, Bohlmann, J, Bousquet, J, Garcia Gil, R, Hvidsten, T. R, de Jong, P, MacKay, J, Morgante, M, Ritland, K, Sundberg, B, Lee Thompson, S, Van de Peer, Y, Andersson, B, Nilsson, O, Ingvarsson, P. K, Lundeberg, J, & Jansson, S. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584.
43. Engel, S. R, Dietrich, F. S, Fisk, D. G, Binkley, G, Balakrishnan, R, Costanzo, M. C, Dwight, S. S, Hitz, B. C, Karra, K, Nash, R. S, Weng, S, Wong, E. D, Lloyd, P, Skrzypek, M. S, Miyasato, S. R, Simison, M, & Cherry, J. M. (2014) The reference genome sequence of *Saccharomyces cerevisiae*: Then and now. *G3: Genes|Genomes|Genetics* 4, 389–398.
44. Ingman, M & Gyllenstein, U. (2006) mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Research* 34, D749–D751.
45. Langmead, B & Salzberg, S. L. (2012) Fast gapped-read alignment with bowtie 2. *Nat Meth* 9, 357–359.
46. Balian, R & Schaeffer, R. (1989) Scale-invariant matter distribution in the universe. *Astron. Astrophys.* 220, 1–29.
47. Gurbatov, S. N, Simdyankin, S. I, Aurell, E, Frisch, U, & Tóth, G. (1997) On the decay of burgers turbulence. *Journal of Fluid Mechanics* 344, 339–374.